

**Human Tissue Pathology Analysis:
Clinically Annotated Tissue Databank Applications
And
Image Compression**

An Honors Thesis

By

Kristen Hampton

Thesis Advisor
Dr. Clare Chatot

Clare L. Chatot

Ball State University
Muncie, Indiana

May 2006

May 6, 2006

Table of Contents

I. Abstract	3
II. Introduction	4
III. CATD	6
a. Introduction	6
b. Background	7
c. Technique	7
d. Impact	11
IV. Compression Study	14
a. Introduction	14
b. Background	14
c. Methods	15
d. Discussion	21
V. Conclusion	23
VI. Acknowledgements	25
VII. References	26

Abstract

An internship at Eli Lilly is an incredible opportunity to be able to experience. There are thousands of different research projects being conducted at Lilly and I was extremely lucky to get to help with two projects. A clinically annotated tissue databank (CATD) was developed about 10 years ago at Lilly, but it still being perfected. CATD is of great importance to the researchers at Lilly because it enables them to find and obtain the necessary tissues they need for the studies they are conducting. During this study, nearly 4,400 patient specimens were entered into CATD making thousands of new tissues available for research purposes. Microscopic sections from over 650 of these tissues were analyzed for cancer types and imaged for inclusion in the database. Also, the specimens from these patients were organized so they could easily be located once requested in CATD. Another project of great significance researched the amount of data loss when an image file is compressed for computer storage. Storage space on hard drives is limited in many medical research companies, which unfortunately limits the number of images researchers can hold on their hard drives. This could potentially even halt some research from occurring. The compression study evaluated compressing images by creating JPEG and JPEG2000 files in color and black and white. Image analysis included calculating how much data was lost at each compression level. There was a very delicate line as to how far an image can be compressed to save hard drive space without losing important data from the image. Overall, JPEG2000 was more efficient at compressing color images without data loss than JPEG. Using liver tissues stained in the H&E, a 10% data loss at a 150 level of compression was acceptable for continued data analysis. Working at Eli Lilly was an amazing experience that gave me a very realistic view of what a research scientist's daily job would be.

Introduction

I am a Pre-Optometry Biology major at Ball State University. During the summer of 2004, I was very lucky to obtain an internship at Eli Lilly. The internship was in the Scientific Imaging Center headed by Michail Esterman. They had chosen me to participate because of my extensive biology background. I also was lucky enough to be able to work extensively in a pathology lab. I was asked to come back for an internship in the same Scientific Imaging Center for the summer of 2005. Two summers at Eli Lilly allowed me to work on many projects and be very involved.

The first project involved work with a clinically annotated tissue databank (CATD). CATD is a tissue databank that has thousand of tissues available for researchers at Eli Lilly. It includes information about the patient that the specimen came from and also any research previously done on the specimen. All of the tissues are either cancer, normal, or normal adjacent to the cancer. I was responsible for cataloging specimens on CATD. The first summer I began the tedious job of entering 1,430 patients into CATD. totaling 4,400 specimens, all acquired from pathologist Dr. George Sandusky. This required much learning and understanding of medical acronyms and definitions in order to correctly enter the patients. Quality Assurance (QA) data was also entered for many of the specimens. The QA data was performed by Dr. Sandusky and it is to make sure that the tissue is correctly identified with percent cancer, necrosis, etc. My first summer I also imaged over 650 of the tissues entered into CATD. The slides were prepared from tissue within a paraffin block and 200X images and if needed 25X images were obtained of the tissues. The second summer another intern, Jennifer Offen, and I began the extensive job of organizing all of Dr. Sandusky's tissues. Dr. Sandusky was the only one that knew exactly where all of the thousands of tissues were located in

his freezers. So, Jennifer and I arranged all 1600 specimen within 8 freezers, 22 shelves, and 290 boxes. Their exact locations were all neatly catalogued in an excel spreadsheet so they could be easily located if requested in CATD.

The second large project which also began the first summer was a compression study. Space to save images is becoming very valuable and many researchers are being limited to the amount of storage space they are allowed. Images are very valuable to many scientific researchers so they really need a way to be able to store them. This is where compressing images would be very useful. There are many different formats for compressing an image. However, in the past in medical studies, researchers have been afraid to compress images because they are fearful they will lose important information with the compression. The objective of this study was to identify how far an image can be compressed using JPEG and JPEG2000 and still be able to extract all the vital data. Images were taken of rat liver sections and a black and white electron microscopy grid to test the levels of compression. The images were originally saved as TIFFs and then resaved as JPEG or JPEG2000 files. Analysis schemes were written by Jeff Hanson to measure the difference between the original TIFF and the JPEG or JPEG2000 files. The difference was then compared to the percent of data a researcher is willing to lose in order to see the level of compression that is possible for their project.

CATD

Introduction:

While working at Eli Lilly, it became obviously clear that human tissue is in high demand for research. The availability can really determine whether a study can continue or not. Since Eli Lilly is a pharmaceutical company and its products are used by humans, it is crucial that relevant studies be done using human tissue. Other animal tissue will work, but more accurate conclusions can be drawn if the study is actually performed on human tissue. Just because an animal tissue produces a certain response does not necessarily correlate to how a human tissue would react in the same situation. Therefore, it is much more applicable to use actual human tissues. Another complication once the human tissue has been obtained, is knowing the background of the tissue. For a few research studies it is acceptable to have tissue without the background information. However, in many studies, it becomes vital to know such information as: age of the patient the tissue was from, race, gender, their personal history, and if appropriate, how they died. For example, for cancer patients, it is very relevant to know the stage of the cancer, how old the patient was, and any cancer in their family history. CATD (Clinically Annotated Tissue Databank) was created to provide a databank of tissues for research, with as much useful information tied to the tissue as possible. Anyone within Eli Lilly can then request and order tissues that appropriately fit their study. The ultimate reason for this project is to implement thousands of new tissues into CATD that all have their patient history and specimen history tied to them. It will make it much easier for researchers to request human tissues that are applicable to the study they are performing

and they will be able to obtain much more information from the study due to knowing the previous history of the tissue.

Background:

CATD was originally started by Eli Lilly and eventually more groups such as Methodist joined in. Any of the organizations could provide tissues to be on CATD. Doctor George Sandusky is a pathologist at Eli Lilly who has literally thousands of different tissues available for his use. There are currently eight freezers for Dr. Sandusky's tissues alone. Dr. Sandusky wanted these tissues to be available through CATD; however, there were many issues before this was possible. There was no systematic inventory of Dr. Sandusky's freezers. He knew exactly where every tissue was, however, this is not very practical since when Dr. Sandusky retires, no one would know specifically where each individual tissue was located. The second issue was that the tissues had a different barcode on the tube than the barcode that was entered on the data sheet. All the barcodes from the excel spreadsheet had been entered into CATD. A sheet would have to be made to correlate between the two barcodes in order to find the correct tissue. Finally, once the barcode had been entered into CATD, the tissue must be able to be found efficiently in the freezers if someone should request it.

Technique:

During my first summer at Eli Lilly, my job was to enter all of Dr. Sandusky's tissues into CATD. The first page of data to enter included personal information about the patient, such as, age, race, birth date, and sex, which was all provided on a spreadsheet. The second page, shown in Figure 1, was a diagnosis of the patient and any medical

CATD - Clinically Annotated Tissue Databank (Rel 3 May 2005 SSL)
 CATD Acct: rx16725 Mike Esteman LY

Current Patient - 255 Site: LY User Logon: rx16725

Patient History 481

Date of Report mm/dd/yyyy: 6/5/06/2006

Pre-existing Conditions: Cancer Head & Neck Larynx ☒

Metastasis Present?: No ☐

Approx. Year of Diagnosis (4 digits): 2001

Description: Invasive moderately differentiated locally metastasizing squamous cell

Submit Patient History Changes

Privacy Statement | Terms of Use | Copyright © 2005 Eli Lilly and Company. All Rights Reserved.
 This site is intended for use by United States residents only.

Figure 1. – Pictured is the web page set-up for the Patient History page. On this page, any previous medical conditions are listed and when they were diagnosed.

problems besides the cancer. This required me to learn many different medical acronyms in order to correctly identify the diagnosis of the tissue and any previous or current health problems with the patient. The third page was the patient's family history, shown in Figure 2, if any had been provided for the tissue. The family history detailed any previous or current cancer within the family and whether it was the same cancer the patient had. The fourth page was the current information about the patient, including, if they were still alive, if so how long had the cancer been in remission and if they died what the cause was. The fifth page, shown in Figure 3, was dedicated to information about the tissue as it was being currently stored. This included the amount of the tissue that remained, whether it was normal, benign, cancer, or normal adjacent tissue, if it was frozen or embedded in paraffin, and the barcode for the specific tissue block. A total of 4,400 specimens were entered into CATD with as much information as was provided for each.

CATD - Clinically Annotated Tissue Databank (Rel 3 May 2005 SSL)
 CATD Acct: rx16725 Mike Esterman LY

Home Request Search Audit Data Define Elements Admin Change Request Logout

Specify Patient Register Patient Information

Current Patient - 235 Site: LY User Logon: rx16725

Family History 219

Family Level:

Date Collected mm/dd/yyyy:

Condition:

Approx. Age Diagnose:

Description:

NOS = Not Otherwise Specified

Figure 2. - Displays the web page layout of Family History for the specimen. The information included on this page includes any type of cancer within the immediate family.

CATD - Clinically Annotated Tissue Databank (Rel 3 May 2005 SSL)
 CATD Acct: rx16725 Mike Esterman LY

Home Request Search Audit Data Define Elements Admin Change Request Logout

Specify Patient Register Patient Information

Current Patient - 235 Site: LY User Logon: rx16725

Specimen 2383

Specimen Barcode:

Specimen Type:

Specimen Method:

Specimen State:

☒ Compliant with Extraction to Freeze for RNA

Amount:

Anatomical Location:

Privacy Statement | Terms of Use | Copyright © 2005 ER Lilly and Company. All rights Reserved.

Figure 3. - This webpage is an example of the Specimen page. It includes information about the exact specimen; such as, weight specimen type, anatomical location, specimen state, and barcode.

Another useful quality of CATD is the quality assurance (QA) page. This page records information from a study by a researcher or pathologist who has looked at a section of the tissue and diagnosed it. This generally was reported as the percent normal, percent cancer, percent necrosis, and the percent adipose tissue. Obviously tissues could have all four, one, or a mixture of these types of tissue. Dr. Sandusky had performed QA data on a large portion of his tissues. Therefore, after entering all the specimen into CATD, QA data for a good portion of them was entered. This required first looking up the tissue in CATD. Once it was located by the barcode, the QA data was entered which generally included the diagnosis assessment and the depth into the tissue the slice was taken from.

Another valuable characteristic of CATD is the ability to put a microscope picture of the tissue slice with the history of the tissue. A block of paraffin with the tissue embedded in it was put into a microtome that cut slices in the depth of microns. The tissue slice would fall into a water bath where it then was taken and put onto a slide. The slides were stained with Hematoxylin & Eosin (H&E) stain to be examined. 100X images of many of the tissue slices were taken on a Leica upright DMRXE microscope. If there was a highly cancerous area surrounded by normal tissue then a 200X image zoomed in on the malignant region was also taken. These images were then saved as TIFF files and imported into CATD. With the image was a diagnosis of the exact slice of tissue. For this, Dr. Sandusky taught me how to recognize cancer, normal, and necrotic cells for any organ tissue type. The histology of a colon cancer tissue is generally structureless masses of cells and many of the crypts appear very elongated with very thick walls. Also, another sign of the cancer in colon cells is the mucin they produce. Not only did I have to recognize cancer in all organ tissue types to accurately take the

image of the section with cancer, but I also then entered the percent tissue type (i.e. percent cancer) with each image. Once the cancer was identified other types also had to be identified, such as, necrosis and adipose tissue. Adipose tissue is primarily fatty tissue. This is very easy to identify because it appears like a lot of balloon inflated cells that are clear. Necrosis is a little harder to identify but not difficult. Necrosis is when the tissue cells have died. This can be identified because generally there is not a defined cell structure.

The final process was to organize all the tissues so they easily could be found by their barcode when requested. Dr. Sandusky's tissues were contained within 8 freezers with 22 shelves. Most of the tissues were in a vial or a bag if it was a whole organ. The first step was to organize the freezers so that all the specimens of the same organ were together. This meant putting all the heart tissue together in a freezer, all the breast tissue together, all the colon tissue together, etc. Once this was completed, we began organizing and cataloging them so they could be retrieved for CATD. At this point, the only identification was the freezer number and the shelf number. It would still take a large amount of time just to find the one vial on a shelf. Boxes were then obtained to put tissues in, and there were approximately 15 boxes on a shelf arranged numerically. As the tissues were being organized into the boxes, an excel spreadsheet was maintained that included the freezer number, shelf, box, the barcode on the vial, and the barcode in CATD it corresponded with. Once a tissue was requested in CATD, the barcode from CATD would be entered into the spreadsheet to find. Once the number was found, the tissue could easily and quickly be located by the information on the spreadsheet. Any new tissues brought in by Dr. Sandusky could easily be entered into the excel spreadsheet.

Impact:

Cataloguing Dr. Sandusky's tissue will make a large impact in his lab. Dr. Sandusky knew where every tissue was in the freezer but it would still take him a fair amount of time to find the exact vial a researcher had requested on CATD. Moreover, it makes sense for someone else in Dr. Sandusky's lab to find the specimens requested because as the pathologist, Dr. Sandusky did not need to spend his time looking for vials. However, no one else in his lab would even know where to start to find the vials. By cataloguing all the vials by a specific location, they could be easily found when needed. All the person must do is enter the barcode from CATD into the excel spreadsheet and once excel found it, the freezer, shelf, box, and true barcode would all be with it. This also makes it a very easy process for when Dr. Sandusky receives new tissues. All that must be done is to go to the freezer containing the correct tissue type (i.e. breast) and find a box it will fit in. If there are no open boxes then a new box would be started and numbered. The location by freezer, shelf, box, and true barcode for the new tissue would then be entered into the excel spreadsheet. This will save a major amount of time that would otherwise be spent searching for tissues. Also, it will save Eli Lilly money because they will not have to hire someone specifically to search for vials that are requested on CATD.

The clinically annotated tissue databank has made long strides for researchers. The 4,400 specimens I entered into CATD are now available to researchers throughout Eli Lilly and other research companies in Indianapolis. Not only is the tissue available to the researchers, but the background information for the patient associated is readily accessible. This has made a large impact on research studies. The researchers can find out much more specific information from a study because they know the information

about the tissue and not just that it is a breast cancer tissue. They will know that it is a stage IV cancer from a Caucasian woman who is 58 years old. They also will know the information that the woman was 45 when the cancer onset and that she has now been in remission for 4 years. Researchers can be much more specific with medicine they are developing because they can know for sure what stage of cancer they are studying. There could be a significantly different treatment for someone in stage IV of breast cancer versus stage I. Also, researchers could perform studies to see if there are differences in specific cancers between sexes, races, and ages. Not only do researchers have access to the background information for the tissue but they can also see an H&E stain of the tissue before ordering it. This way the researchers can make their own pathological assessment of the tissue to reassure it is the tissue they need for their study. Finally, there is Quality Assurance data and research pages that provide results of tests previously performed on the tissue. Having a record of the previous research and results on a tissue specimen can help prevent unnecessary duplication of repetitive tests or analyses. This would save much needed time and money.

Compression Study

Introduction:

Images represent large data sets that are more complicated to manage than a typical file like an excel spreadsheet. The complications come from different formats, different dimensions, and size. The topic of this study is dealing with size by using compression.

When compressing images at high levels data is lost. So that the results of a research project will not be compromised, we need to be able to determine how much data is lost and what level of data loss would be acceptable. Common compression techniques are LZW (Lempel Ziv) encoding, jpeg compression, jpeg2000, and zip files. LZW and zip files typically don't save more than 2-3% of space making them unacceptable as a space saving tool, therefore, they were not researched during this project. This study will be looking closely at jpeg compression and jpeg2000 to assess the data lost relative to the amount of compression.

There are three key questions that need to be addressed before accepting compression of scientific images. At what level of compression can you still extract the data needed from an image? How subtle a change can you detect in an image at a compression level? Can you balance storage cost against compression-induced quality loss? Our aim is to determine a metric that can help recommend an acceptable compression technique and quality level to maintain usability of the image for data analysis.

Background:

There are several topics that repeatedly come up when discussing image quality and image compression. Many images in the medical field are saved as TIFFs, tagged

image file format, because there is no loss of data, however, there is also no compression. This has begun to cause a problem for saving images because TIFF images take up such a large amount of disk or hard drive space when stored. Compressing images can be a relief if there is not infinite storage space. JPEG is a very common compression format for photographic images which “uses an 8x8 block size discrete cosine transform (DCT)” (Wikipedia). The more recent format of compression is JPEG2000. It is a “wavelet-based image compression standard” that was created with the aim of improving the original JPEG.

A paper very similar to the study is “Image Compression in Morphometry Studies Requiring 21 CFR Part 11 Compliance: Procedure is Key with TIFFS and Various JPEG Compression Strengths” by Mark Tengowski, which evaluates the compression of JPEG and the amount of data lost. He concluded that the level of the compression with meaningful data still able to be extracted depended highly on the object of interest. A black and white image of a grid can be compressed much farther than a color image of a liver tissue and still obtain the data needed (Tengowski 260). The purpose of our study is to support Mark Tengowski’s research and take it a step farther since JPEG2000 was not available when Tengowski’s research was concluded.

Methods:

Images were captured on a Leica upright DMRXE microscope. A 200X PlanApo objective (NA=0.6) was used on the microscope, so the total magnification was 200X for each. A SPOT RT color camera was mounted on the microscope and used to capture images to a PC running Windows 2000. There were four different sample types studied. The first was a 400-mesh electron microscopy grid. The grid provided a very uniform target for image analysis. The second sample was an H&E stained rat liver section. The

liver acted as a good typical pathological sample very common in the medical field.

Third we had a set of fluorescently tagged cells. The final set of samples was a batch of electron micrograph images sent to be analyzed by a pathologist. However, only the study on the electron microscopy grid and rat liver will be discussed within the paper.

Images were captured of each sample type at 200X and saved as a TIFF file. Each image was then converted into both JPEG and JPEG2000 files by choosing a compression level. The JPEG images were compressed in Image-Pro Plus and the JPEG2000 images were compressed in Photoshop with a LuraWave plug in. The images were then converted back into TIFFs for analysis in Image-Pro Plus v5.1.0.20. A variety of JPEG and JPEG2000 compression levels were assessed.

The image-processing analysis routines were created by Jeff Hanson, an IT at Eli Lilly. The analysis compared the compressed image to the original TIF image to assess the amount of data lost. The paper by Mark Tengowski suggested techniques for analyzing grid and liver samples. We adapted these techniques to the analysis of our grid and liver images. The liver analysis was of an H&E stained rat liver tissue which was segmented into three regions to be measured: the nuclei, sinusoidal space, and cytoplasm regions. The nucleus is a single membrane-bound organelle of a cell which stores the genetic information and controls chemical reactions (Wikipedia). The nucleus, when stained with H&E, generally stains a dark blue. Sinusoids are small blood vessels within the liver (Wikipedia). When stained with H&E, sinusoids appear white. Finally, the cytoplasm is a homogenous substance that fills the cell (Wikipedia). It generally gives the cell its support and rigid shape. When stained with H&E, the cytoplasm stains pink. The grid analysis used a routine that is based on brightness or darkness of the object. The measurements were the maximum, the minimum, the mean, and the range of values for

perimeter length. Fluorescence analysis was computed by the root mean square and root mean cube (RMS/RMC) for each image. Each cell was measured for mean intensity and standard deviation of intensity. The electron micrographs used a scientific evaluation of each image compression level for usability. All of the data from the analysis's were transferred into an excel sheet for more calculations to be performed.

Analysis:

JPEG worked very well for storage of black and white images. The way that JPEG compresses images, however, does not work very well for color images.

A. JPEG

B. JPEG2000

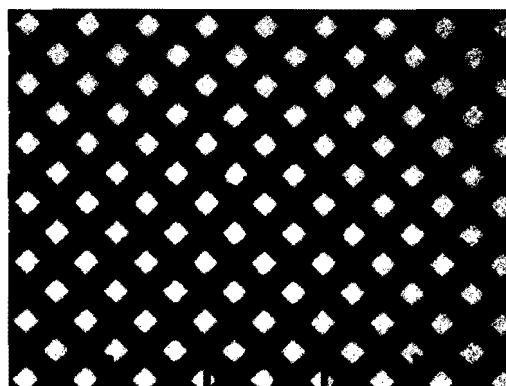
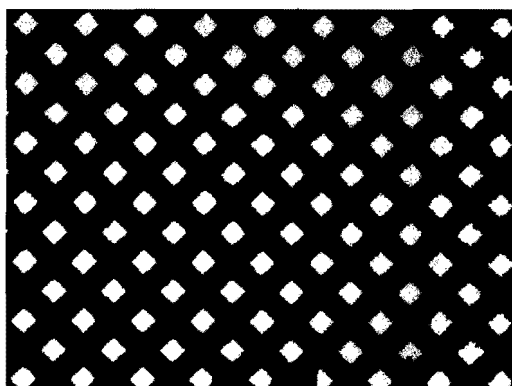


Figure 1. – 400-mesh electron microscopy grid images compressed by both JPEG and JPEG2000. The images were taken on a Leica upright DMRXE microscope magnified by 200 times. Image A is compressed with JPEG. The far left section is the original TIFF image. As the image progresses to the right, the tick marks indicate a higher level of compression for each section. Image B is compressed with JPEG2000. The image format follows the same set-up as the JPEG image. Both images show acceptable compression for each level and still contain the data needed.

The 400-mesh electron microscopy grid in Figure 1 was compressed at multiple levels and re-assembled into a single image. The far left is uncompressed while the far right represents the maximum compression. In each image, the farthest section to the left is the original TIFF image. As the image progresses to the right, each tick marks a higher level of compression by either JPEG or JPEG2000. For both JPEG, image A, and JPEG2000, image B, the image structure is maintained, even at the highest compression

levels. At all compression levels white versus black and the lines between the two are still very clear. At the highest level of compression the borders begin to become fuzzy, however, the data extracted is still very close to the original.

A. JPEG

B. JPEG2000



Figure 2.– Rat liver images compressed by both JPEG and JPEG2000. The images were taken on a Leica upright DMRXE microscope magnified by 200 times. Image A is compressed with JPEG. The far left section is the original TIFF image. As the image progresses to the right, the tick marks indicate a higher level of compression for each section. Image B is compressed with JPEG2000. The image format follows the same set-up as the JPEG image. Both JPEG and JPEG2000 become unacceptable at a certain point. They both become very blurry and JPEG even loses color at its highest compression level.

An H&E slide of rat liver, shown in Figure 2, was compressed at multiple levels and re-assembled into a single image. As before, the far left section of each image is the original TIFF. As the image progresses to the right, each tick mark represents a higher level of compression by either JPEG or JPEG2000. The farthest section to the right of each image is the maximum level of compression for both JPEG and JPEG2000. However, these levels of compression are not the same for both. JPEG2000 compresses to a much smaller level than JPEG. As seen in Figure 3, JPEG2000 compresses to the level of about 650 times smaller than the TIFF file, however, JPEG only compresses to about 200 of its original size. The left of each image is uncompressed while the far right represents the maximum compression. The JPEG, image A, compression becomes blurrier faster than JPEG2000, image B. The JPEG compression also loses significant

color information at high compression levels. By contrast, the JPEG2000 is still recognizable at the highest compression levels, even though it is very blurry. Even at the highest level of compression for JPEG2000 the nuclei, sinusoidal area, and cytoplasm can all still be distinguished.

Most medical images are in color so it is important to look closer at the H&E image of the rat liver; it is the real-world example. It is easy to tell merely from looking at the compressed images that JPEG does not do as well as JPEG2000, and eventually even loses color.

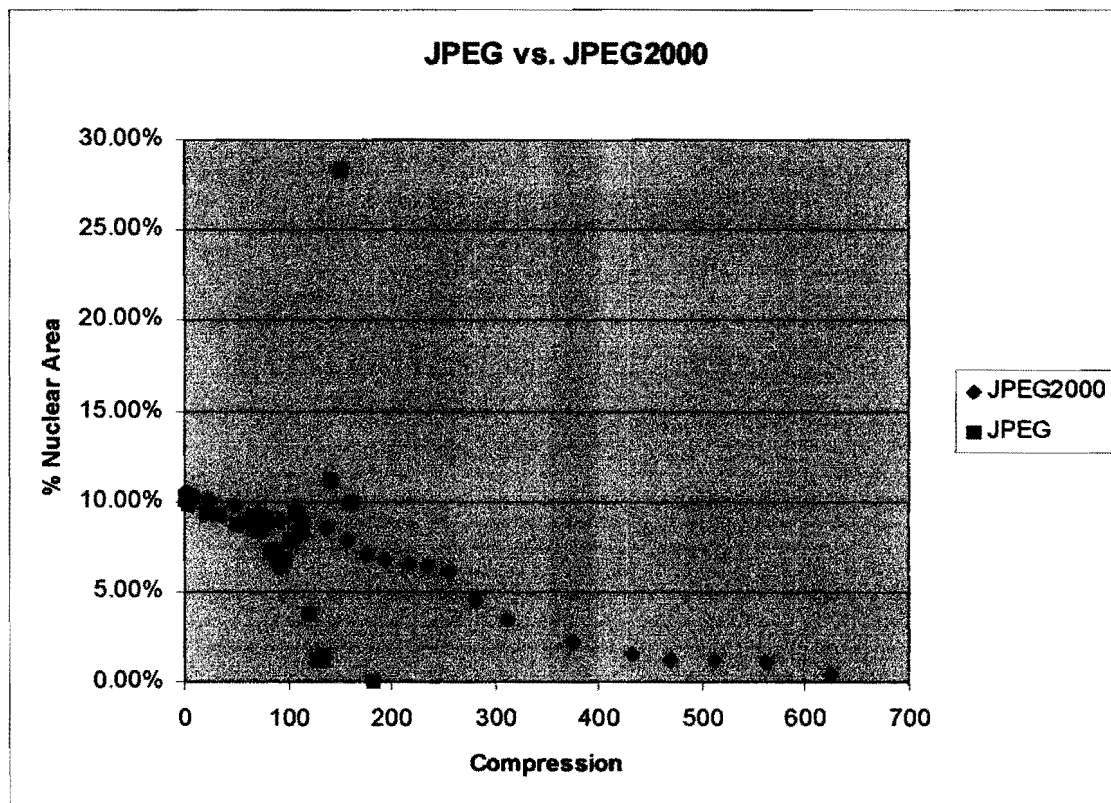


Figure 3.-This graph represents the amount of nuclear area at several levels of compression for both JPEG and JPEG2000. Zero is the original TIFF image and 700 is the highest level of compression acknowledged. JPEG only maintains acceptable compression up to about 75, after this JPEG is very unpredictable as to the accuracy of the data. JPEG2000 performs slightly better, maintaining acceptable compression to about the 150 compression level. After the 150 compression level, JPEG2000 begins to slowly decline in its accuracy.

Figure 3 shows the percent Nuclear Area for different compression levels in both JPEG and JPEG2000. JPEG2000 has a slow gradual decline in data loss. However, you can see from the chart that JPEG fails at a much lower compression for the data. JPEG can only be accepted up to the 75 compression level, where as, JPEG2000 perform slightly better maintaining acceptance up to the 150 compression level.

Once it was determined that JPEG2000 compresses color images better than JPEG, it was important to see where the data was lost when compressed by JPEG2000. There were three main items that were analyzed in the rat liver, each being altered by the amount of compression.

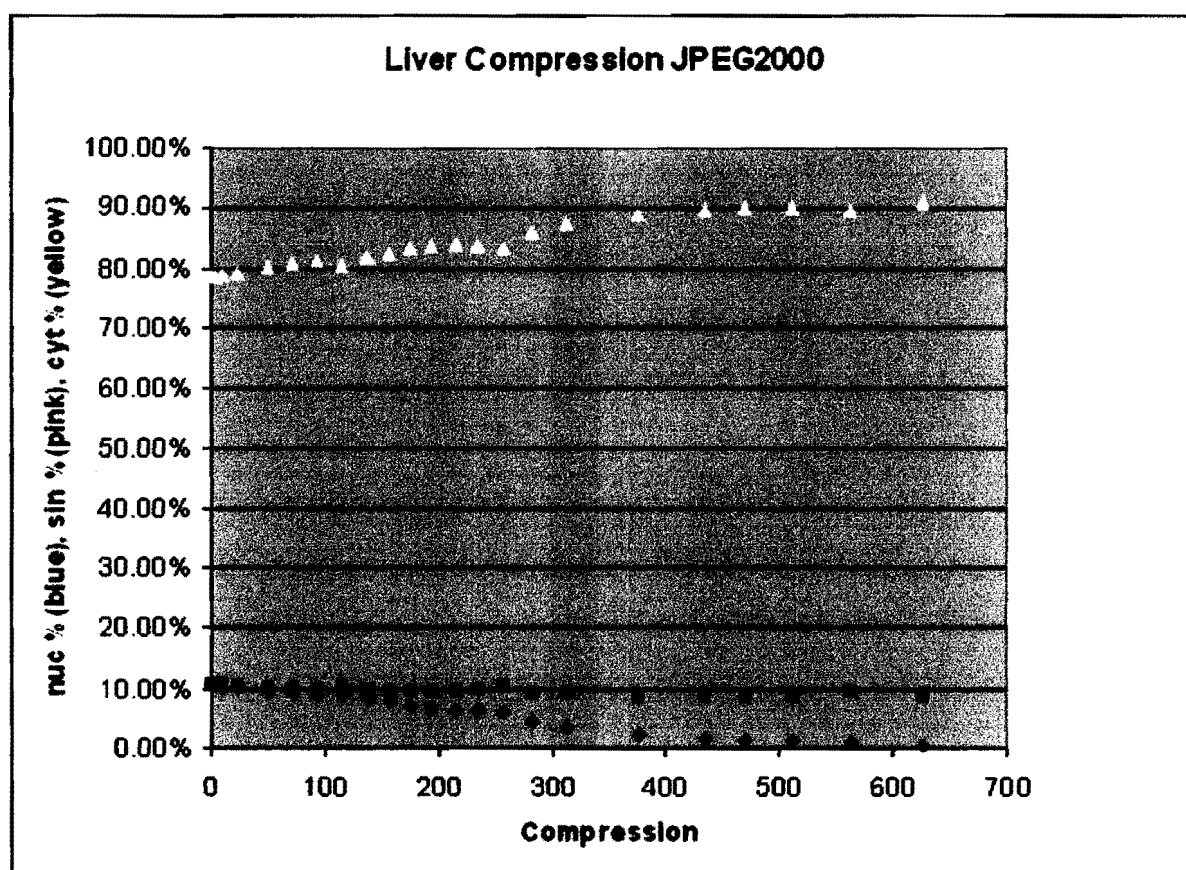


Figure 4.- The graph represents the nuclear area (nuc), sinusoidal area (sin), and cytoplasm (cyt) present at each compression level for JPEG2000. At about the 150 compression level JPEG2000 begins to lose its accuracy. The nuclei and the sinusoidal space both decrease in area, where as, the cytoplasm increases in area.

The three items measured in the liver H&E were the nuclear area, sinusoid area, and cytoplasmic area. Figure 4 shows how each area changes with each compression. The nuclei area decreased in area with increasing compression starting around a compression level of 150, while at the same time, the cytoplasm increased in area around the same compression level. The more the image is compressed the smaller the nuclei become which leads to more cytoplasmic area. However, the sinusoidal area really seems to maintain constant even to the maximum level of compression and only varies slightly starting around the 250 compression level.

Discussion:

The farther an image is compressed the more data loss there is. However, it is also true that the more compressed an image is the more space that is saved. To be able to find an acceptable compression level the researcher must decide how much of a data loss is acceptable. For some fields of study this might be greater than others. For example, a ten percent data loss here is what compression level you can go to for good liver images and to be able to extract the data necessary. A ten percent data loss can be correlated to around a 150 level of compression for JPEG2000. However, with the grid, the data is still very relevant at 999 compression, the highest level of compression. It must be kept in mind though, that is completely depends on how much data loss the researcher is willing to accept. In all scenarios, JPEG2000 is best for compressing color images and having the least data loss. This is due to the procedure used to compress the images. However, with black and white images JPEG performs just as well as JPEG2000. The subject of the image must be analyzed along with the acceptable data loss before there can be a suggested compression level and format for compression. The potential of this study and subsequent studies is to be able to have a series of steps to

evaluate and recommend if compression will be useful and what compression level data can be accepted without compromise.

Conclusion

The work that I did at Eli Lilly has significant impact. There are many people impacted by the CATD study alone. CATD is very useful for many researchers at Eli Lilly. They can obtain tissues for their study and know all of the background information for the tissue. It also allows them to be much more specific with their study by knowing the information. They could do a study on the prevalence of breast cancer in smokers to see if there is a correlation. They could even go as far as to determine correlations between environmental variables and the stages of the breast cancer, given the patient and specimen history. Also, with the QA data entered, they can be positive they are ordering exactly what the tissue was recorded as. The work Jennifer Offen and I completed impacts Dr. Sandusky's lab in significant ways. By arranging and organizing his specimens, anyone in his lab is now able to find an exact specimen. Also, when a specimen is requested within CATD, hours will not be spent searching for the exact specimen. There are much more imperative things in their lab they need to be working on than searching for a tissue. The excel spreadsheet allows them to enter the barcode number and it will immediately show the exact location of the tissue.

Many researchers, and possibly someday a much more broad group of people, will be affected by the compression study. Storage space has really become an issue at many medical research companies. This data obtained from the compression study will allow researchers to save storage space without losing important information from the image. Unfortunately it is not a strict black and white line as to the compression level that should be used to save storage space and maintain important data. It is difficult because there are many factors that affect this, such as, if the image is black and white or color, the actual subject of the image, and the amount of data loss the researcher is

willing to allow. However, eventually there will be a guide describing all of these variables that will allow them to know the compression level to use.

I worked on many different projects while I was at Eli Lilly. Each one of them showed me a different side of medical research and what all goes into an actual research study, including planning and even obtaining the resources to do the study. It is very different from an undergraduate biology experience in a classroom laboratory because you do not have to worry about the planning or obtaining the materials, the professor simply tells you what to do. The internship gave me a very realistic idea of how many research biologists and chemists work every day. By working in the Scientific Imaging Center, I interacted with many different researchers who all needed parts of their research imaged. I also interacted with many scientists that helped me with my projects because they all have their own expertise. It also helped me form many connections not only within Lilly but outside of Lilly. It gave me the opportunity to job shadow a surgeon, Dr. Frank Lloyd, at Methodist Hospital, who also sat on the Admissions Board of IU School of Medicine. This was very important at that time, since I was planning on applying for medical school. Finally, I learned every little detail about any microscope available, including uprights, inverted, and confocal microscopes. It was a wonderful opportunity that gave me a very realistic experience of the work life of a scientist.

Acknowledgements

- I would like to thank Dr. Clare Chatot for being my Honors Thesis advisor. She was a wonderful advisor and a tremendous help in writing the paper.
- I would also like to extend a special thanks to Michail Esterman and Jeff Hanson at Eli Lilly. My experience as an intern at Eli Lilly and the opportunity to work on these projects was completely thanks to them. They were wonderful mentors.
- Finally, I would like to thank Dr. George Sandusky and the researchers in his lab. Dr. Sandusky shared his expertise and wealth of knowledge with me and always had time for my thousands of questions referring to histology.

References

- “Aperio Compression performance measurements”. 22 Feb 2005. Aperio Technologies, Inc. 26 July 2005
http://www.aperio.com/documents/Aperio_Compression_performance_measurements.pdf.
- Tengowski, Mark W. (2004). Image Compression in Morphometry Studies Requiring 21 CFR Part 11 Compliance: Procedure is Key with TIFFS and Various JPEG Compression Strengths. *Toxicologic Pathology* 32, 258-263.
- J.C. Russ (2002). Image Processing Handbook, Fourth Edition, CRC Press, Boca Rotan, FL.
- Wikipedia the free encyclopedia*. July 2005. Wikimedia Foundation, Inc. 20 July 2005
<http://en.wikipedia.org>.